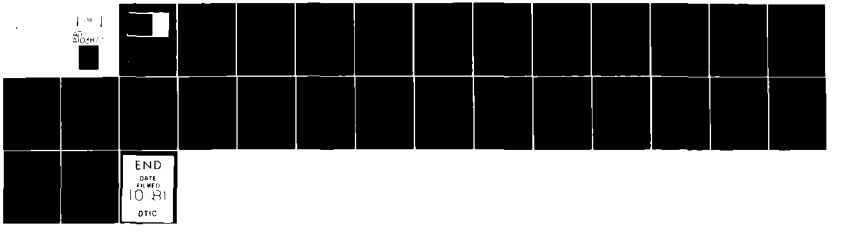
AD-A103 877 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
THE RELATIONSHIP BETWEEN THE NUMBER OF FACTORS AND SIZE OF AN E--ETC(U)
JUL 81 D V LINDLEY DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2246 NL



END
DATE FILMED
10-81
DTIC

MRC Technical Summary Report # 2246

THE RELATIONSHIP BETWEEN THE NUMBER OF FACTORS AND SIZE OF AN EXPERIMENT

D. V. Lindley

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

July 1981

(Received May 7, 1981)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

THE RELATIONSHIP BETWEEN THE NUMBER OF FACTORS AND SIZE OF AN EXPERIMENT

D. V. Lindley

Technical Summary Report #2246

July 1981

ABSTRACT

If the number of factors in an experiment is increased, does it necessarily follow that the size of the experiment must increase to achieve a satisfactory analysis? In some common situations the answer is No. The present paper discusses a model suggested by medical diagnostic problems in which the answer is Yes: indeed, the increase in size is exponentially fast. The conclusion is drawn that statisticians should be cautious before embarking on the study of data with large numbers of factors because the data may be inadequate for a sensible analysis. The basic, mathematical tool is the Kullback-Leibler number which measures the discrimination between the possibilities. Calculation of these numbers uses interactions, forming a basis for all the effects that might occur.

AMS (MOS) Subject Classifications: 62K99, 62G99, 94B99

Key Words: diagnosis; error-correcting codes; errors in variables; experimental design; interactions; Kullback-Leibler numbers; non-parametric methods.

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Suppose that on each of a number n of experimental units the values of m factors, or variables, are measured: for example, n patients may be tested for the presence of m symptoms. Then it seems intuitively reasonable that as the number of factors increases it may be necessary to observe more units in order to understand the influence that the factors have. In many situations studied in statistics this is not so. In the present paper a contrary case is considered in which the growth in n needs to be exponentially fast in m; for example each extra factor may mean a 25% increase in the number of units needed. The case is suggested by medical diagnostic problems, though as a description of the medical reality it has a number of defects. Nevertheless the general moral that increasing complexity of factors means more data is surely true there.

The practical import of the results is that scientists and their statistical advisors should be wary of handling data sets with a large number of factors simply because n may not be large enough to permit reliable conclusions. The tendency these days to try to make sense of a lot of data - a tendency made possible by the advent of powerful computers - should sometimes be resisted. More thought should go into the sensible design of experiments, and the statistician's role should not be confined to data analysis, which may be a hopeless task.

The mathematical analysis uses a concept of distance between the various possibilities that might explain the data. This distance is called a Kullback-Leibler number. In the model studied the number of possible descriptions gets to be so large that they crowd together even in the wide-open regions of m-space and hence become close together and difficult to separate. In other situations this crowding does not occur. The mathematical considerations are closely related to those that arise in the study of error-correcting codes: for if two messages are too close they also cannot be separated.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

THE RELATIONSHIP BETWEEN THE NUMBER OF FACTORS AND SIZE OF AN EXPERIMENT

D. V. Lindley

1. Introduction

With the growth in computing power statisticians have turned their attention to data sets involving a large number of factors, or variables. In this paper we consider the relationship between n, the size of the data set,or experiment, and m, the number of factors. Intuitively it appears plausible that as the number of factors increases so should the size of the experiment in order to unravel the increasing complexity that could arise with many factors. The following example shows that this is not necessarily so.

In an experiment in which all factors are at two levels and a complete, factorial arrangement is used, any contrast is a mean of 2^{m-1} values against a similar mean and has variance $2^{-(m-2)}$ times that of any unit. Hence the variance of any contrast per. unit of experimentation is 4, irrespective of m, and the information gained about the contrast per. unit is the same irrepective of the number of factors involved. In this sense the number of factors does not influence the need for more data.

Another example arises with the multivariate normal distribution with n observations on each of m variables. Each mean has n values to estimate it, as has each variance. Equally each covariance

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

has n pairs of values available. Again the number of variables might appear to have no effect. A case with m much larger than n is worth contemplating. This case is not so clear-cut since standard multivariate theory has difficulties whenever n is less than m.

In this paper we study a model and show that there n does have to increase with m; indeed the increase is exponentially fast in order that the many possibilities introduced by an increase in m can be investigated adequately. The model was suggested by problems of medical diagnosis where upwards of 40 symptoms (factors) may be observed on a patient and it is required to diagnose the disease from these. Typically the presence or absence of a symptom is not too well-defined and the model allows for errors in this regard. We do not pretend to have made a contribution to actual diagnosis but merely to indicate the difficulties that might arise when many symptoms are studied.

2. The Model

All quantities are binary, taking the values plus or minus one, or simply + or -. There are m binary factors $\xi = (\xi_1, \xi_2, \ldots \xi_m)$ and dependent on them is a binary response η described by a response function $\eta = \delta(\xi)$. For each unit in the data set every ξ_i has chance 1/2 of being + independently of the other factors: thus each ξ has chance 2^{-m}. Each of the factors and the response for a unit are observed with a possible error. ξ_i is observed as x_i and η as y with $p(x_i \neq \xi_i) = p(y \neq \eta) = p$, $p < \frac{1}{2}$. These are independent and

independent of the factors or response. We write $x = (x_1, x_2, \ldots x_m)$. On the basis of observations $z = (x,y)$ on each of n units it is required to make inferences about δ, the response mechanism. A few comments on the model are now offered.

The structure through ξ's and η measured with error seems reasonable for many situations (as medical diagnosis) where the view is that were the correct quantities to be measured the response would be determined. It therefore differs from the factorial situation mentioned above in that there are errors in the factors (economists speak of errors-in-variables models). Where the model is seriously deficient as a description of reality is in the chance structure imposed for simplicity in the subsequent analysis. It might be possible to have ξ_i (and hence x_i) equally likely to be + or - by defining the factors suitably: for example, a continuous variable could be dichotomized at the median. But it is unrealistic to assume that the factors are independent. Our reason for doing so is that we do not know of a simple way of describing adequately dependence amongst binary factors: indeed, this is a major problem in a satisfactory statistical treatment of diagnosis. Another defect is that the errors have the same chances for each factor. In the medical situation this is not true: age is more accurately determined than blood pressure. Different chances could be studied with the penalty of a substantial increase in algebraic and computational complexity. Our defense of the model is that our interest lies primarily in the relationship between m and n as far as determining δ is concerned, and that the general form of

this relationship might not be too disturbed by more realistic modelling of the chances. The exact numerical values calculated below should not be taken seriously: but we hope that their orders of magnitude might be.

It might be suspected that problems would arise as m increases since the number of δ's is 2^{2^m}, making the determination of δ more difficult. Notice that the model has no parameters and the argument is essentially non-parametric.

3. Discrimination between response functions:

Let δ_1, δ_2 be any two different response functions carrying ξ into η and at some stage of the experiment let $p(\delta_1)$ denote the probability attached to δ_1. If an additional unit is observed to have $z = (x,y)$, then the log-odds for δ_1 against δ_2 will become

$$\log \frac{p(\delta_1|z)}{p(\delta_2|z)} = \log \frac{p(z|\delta_1)}{p(z|\delta_2)} + \log \frac{p(\delta_1)}{p(\delta_2)}$$

by Bayes theorem. Conditional on δ_1, the expected change in log-odds will be

$$\Delta(\delta_1 : \delta_2) = \sum_z p(z|\delta_1) \cdot \log \frac{p(z|\delta_1)}{p(z|\delta_2)} \qquad (3.1)$$

a Kullback-Leibler number. (Notice that in the notation the first argument of Δ is the δ for which the expectation is being calculated - the "true" value - whereas the second argument is the alternative value. In general $\Delta(\delta_1 : \delta_2) \neq \Delta(\delta_2 : \delta_1)$. We sometimes abbreviate to Δ_{12}.)

As more units are added the log-odds change by the addition of the appropriate log-likelihood-ratio, and the expected change for n units is $n\Delta_{12}$. To achieve a prescribed level of discrimination between δ_1 and δ_2 when δ_1 obtains we would need the log-odds to attain a prescribed level R and expect to see n units to achieve this where $R = n\Delta_{12}$. Hence the sample size required is proportional to Δ_{12}^{-1}. For a given δ_1 the discrimination is most difficult and requires the largest sample size for the alternative δ_2 that minimizes $\Delta(\delta_1 : \delta)$ over all δ not equal to δ_1. Since δ_1 is itself initially unknown, the size of experiment is related to the least Kullback-Leibler number $\Delta(\delta_1 : \delta_2)$ over all unequal δ_1, δ_2.

In the discussion that follows it will be supposed that all δ's are initially equally likely so that the original log-odds are zero. If this is not so then the appropriate log-odds should be added to $n\Delta_{12}$ before the evaluation.

The Kullback-Leibler numbers are found by first calculating

$$E_{12} = E(\delta_1 : \delta_2) = \sum_z p(z|\delta_1) \cdot \log p(z|\delta_2) \qquad (3.2)$$

for all δ_1, δ_2 including $\delta_1 = \delta_2$; then $\Delta_{12} = E_{11} - E_{12}$.

4. Interactions

The evaluation of the Δ's for general δ presents difficulties so we consider first an important subclass of δ's called interactions. In applications interest might centre on δ's that depend only on a subclass of the m factors. If changing ξ_i has no effect on δ whatever be the values of the other factors we say ξ_i is <u>irrelevant</u> to δ: the remaining factors are called <u>relevant</u> to δ. The set of relevant factors for δ is denoted $R(\delta)$.

δ is an <u>odd (even), k-factor interaction</u> if $R(\delta)$ contains k factors and $\eta = 1$ iff an odd (even) number of these factors are $+$. The odd and even interactions with the same $R(\delta)$ are called <u>complementary</u>. The 1-factor interactions may be called the <u>main effects</u> of the single, relevant factor. To complete the situation we need the two complementary zero-factor interactions in which $\eta = -(+)$ for all ξ. Notice that interactions, although similar to interactions in factorial experiments, are different in that they occur in complementary pairs, the one being obtained from the other by interchanging $+$ and $-$ in the response values. All the odd interactions for 3 factors are listed in the Table.

In an abuse of notation, a k-factor interaction will often be denoted δ_k. Two different k-factor interactions will be written δ_k and δ_k'.

The reader not interested in the proofs may at this point proceed directly to Theorem 2 below: the I- and J-functions are defined immediately after the statements of lemmas 2 and 3. The mathematical techniques employed are related to those used in error-correcting codes (for example, Lin (1970)) but the development given here is self-contained.

5. Probability evaluations

For any $\zeta = (\xi, \eta)$ either $\eta = \delta(\xi)$ or not. The 2^m ζ's for which $\eta = \delta(\xi)$ are said to <u>agree</u> with δ and the set of them is denoted $Z(\delta)$. Since all 2^m ξ's have the same chance

$$p(z|\delta) = 2^{-m} \sum_{\zeta \in Z(\delta)} p(z|\zeta) \tag{5.1}$$

and

$$p(z|\zeta) = p^s q^{m+1-s}$$

where $q = 1-p$ and s is the number disagreements between z and ζ, a disagreement being where $\xi_i \neq x_i$ or where $\eta \neq y$. We refer to s as the (Hamming) <u>distance</u> between z and ζ.

The following simple result will be used repeatedly in the argument. If s_1 is the number of disagreements in respect of relevants factors and the response, and s_2 the number of disagreements for irrelevant factors, so that $s = s_1 + s_2$, we may write

$$p(z|\zeta) = p^{s_1} q^{k+1-s_1} \cdot p^{s_2} q^{m-k-s_2} \tag{5.2}$$

(s_1 is called the relevant distance). In summing $p(z|\zeta)$ over $\zeta \in Z(\delta)$ all combinations of the irrelevant factors will occur since they have no effect on η. Hence the second factor in (5.2) will yield $(p+q)^{m-k} = 1$. Consequently in evaluating (5.1) it is enough to consider only the relevant factors and the response. This is termed the relevancy principle.

We proceed to the evaluation of $p(z|\delta)$ - equation (5.1) - when δ is a k-factor interaction. There are two cases to consider according as $z \in Z(\delta)$ or not. If $z \in Z(\delta)$ there is one ζ identical with z and distant 0. There are no ζ's distant one since a disagreement for a ξ_i will mean a disagreement either for another factor or for the response by the interaction property. There are two types of ζ distant 2: those with $\eta = y$ and two disagreements in relevant factors; and those with $\eta \neq y$ and one disagreement in relevant factors. There are $\binom{k}{2}$ of the first type and k of the second, a total of $\binom{k+1}{2}$. (The other factors are omitted by the relevancy principle.) Continuing with this line of reasoning we find that if $z \in Z(\delta)$ there are $\binom{k+1}{s_1}$ ξ's at relevant distance s_1, provided s_1 is even, and none if s_1 is odd. If $z \notin Z(\delta)$ the same result holds with odd and even interchanged. Hence

$$p(z|\delta_k) = 2^{-m} \sum \binom{k+1}{s_1} p^{s_1} q^{k+1-s_1}$$

where the summation is over s_1 even (odd) according as $z \in (\notin)\ Z(\delta_k)$; $0 \leq s_1 \leq k+1$.

Lemma 1. In n Bernoulli trials with chance p of success, the chance of an odd number of successes is

$$\alpha_n = \frac{1}{2} - \frac{1}{2}(q-p)^n \quad .$$

This well-known result is easily established by writing down and solving a recurrence relation for α_n. Notice that since $p < \frac{1}{2}$, $p < q$ and $\alpha_n < \frac{1}{2}$: also α_n increases with n. Direct use of the lemma gives

Theorem 1. For a k-factor interaction δ_k

$$p(z|\delta_k) = \begin{cases} 2^{-(m+1)}\{1 + (q-p)^{k+1}\} = 2^{-m}(1 - \alpha_{k+1}) & \text{if } z \in Z(\delta_k) \\ 2^{-(m+1)}\{1 - (q-p)^{k+1}\} = 2^{-m}\alpha_{k+1} & \text{if } z \notin Z(\delta_k) \quad . \end{cases}$$

6. The Kullback-Leibler numbers

Lemma 2. For a k-factor interaction δ_k

$$E(\delta_k : \delta_k) = I(\alpha_{k+1}) - m \log 2 \quad .$$

Here $I(x) = x \log x + (1-x) \log(1-x)$, Shannon's information function. From Theorem 1 and (3.2) and the consideration that both $Z(\delta_k)$ and its complement contain 2^m members

$$E(\delta_k : \delta_k) = 2^m \cdot 2^{-m}(1 - \alpha_{k+1}) \log 2^{-m}(1 - \alpha_{k+1})$$

$$+ 2^m \cdot 2^{-m}\alpha_{k+1} \log 2^{-m}\alpha_{k+1}$$

which easily reduces to the expression given.

Lemma 3. If δ_j and δ_k are respectively j- and k- factor interactions which are neither the same nor complementary

$$E(\delta_j : \delta_k) = J(\alpha_{k+1}) - m \log 2 \quad .$$

Here $J(x) = \frac{1}{2} \log x(1-x)$. $\log p(z|\delta_k)$ takes only two values (Theorem 1) and the multiplier of $\log 2^{-m}(1-\alpha_{k+1})$ will be $\sum_{z \in Z(\delta_k)} p(z|\delta_j)$. Each term in this summation will be either $2^{-m}(1-\alpha_{j+1})$ or $2^{-m}\alpha_{j+1}$ according as $z \in Z(\delta_j)$ or not. But since δ_j and δ_k are interactions (neither equal nor complementary) both $Z(\delta_j) \cap Z(\delta_k)$ and $\bar{Z}(\delta_j) \cap Z(\delta_k)$ contain 2^{m-1} members. Hence the multiplier is

$$2^{m-1} \cdot 2^{-m}\alpha_{j+1} + 2^{m-1} \cdot 2^{-m}(1-\alpha_{j+1}) = \frac{1}{2} \quad .$$

Similar considerations apply to the multiplier of $\log 2^{-m}\alpha_{k+1}$ and the result is immediate.

Theorem 2. If δ_j and δ_k are respectively j- and k- factor interactions which are neither the same nor complementary

$$\Delta(\delta_j : \delta_k) = I(\alpha_{j+1}) - J(\alpha_{k+1}) \quad .$$

If they are complementary $(\delta_k = \delta_j')$

$$\Delta(\delta_j : \delta_j') = 2\{I(\alpha_{j+1}) - J(\alpha_{j+1})\}$$

(if $\delta_j = \delta_k$, the Kullback-Leibler number is clearly zero.)

The theorem is obvious from lemmas 2 and 3 since $\Delta_{12} = E_{11} - E_{12}$. The special, complementary case follows easily by following an argument parallel to that used to establish lemma 3.

7. Expected size of experiment

The implications of Theorem 2 are best understood by referring to the Figure which sketches the functions $I(x)$ and $J(x)$ in $0 \leq x \leq \frac{1}{2}$. Properties of these functions that are cited are all easily established by use of the differential calculus and the proofs are accordingly omitted. $I(x)$ is decreasing and $J(x)$ increasing in the interval and they are equal to $-\log 2$ at $x = \frac{1}{2}$. Since α_j increases with j, $\Delta(\delta_j : \delta_k)$, which is of course positive, decreases both with j and with k: that is, the higher the order of the interactions (or equivalently, the more factors involved in either the true or alternative response) the more difficult it is to separate them. As we saw in section 3, the expected size of experiment is inversely proportional to the Kullback-Leibler number and with m factors the least number will be $\Delta(\delta_{m-1} : \delta'_{m-1})$ in comparing one (m-1)-factor interaction with another. (There are only 2 complementary m-factor interactions which are more easily distinguished because of the doubling of the Kullback-Leibler number that then occurs: see Theorem 2.)

Corollary 1. In the assumed model with m factors and attention confined solely to interactions, the expected size of experiment to distinguish adequately between all response functions is inversely proportional to

$$\Delta(\delta_{m-1}: \delta'_{m-1}) = I(\alpha_m) - J(\alpha_m) = \frac{1}{2}(q-p)^m \cdot \log \frac{1+(q-p)^m}{1-(q-p)^m} .$$

The behaviour of this function with m can be appreciated by noting that for large m it is approximately $(q-p)^m$, so that the increase of n with m is about exponentially fast. The following values for $p = 0.1$ underline this point:

m	2	4	8	16	32
$\Delta_{m-1,m-1}$	.4852	.1782	.0284	7.925×10^{-4}	6.277×10^{-7}

so that an experiment with 32 factors (or medical symptoms) can be expected to require about 10^6 times more observations than one with only 2 factors. As explained in section 2 these numbrs are not to be taken too seriously, they can only express orders of magnitude. Notice that they are exaggerated by the fact that the worst case is being discussed. With a distribution over δ's the sizes would fall. For example, if one was sure that only k out of the m factors were relevant (but not which k) then $\Delta_{k,k}$ not $\Delta_{m-1,m-1}$, would indicate the order. Against this, only interactions are being considered, 2^m out of the 2^{2^m} possible functions, so that even larger sizes are possible: see section 8.

If $j > k$ in Theorem 2, $\Delta(\delta_j : \delta_k) > \Delta(\delta_k : \delta_j)$. Taking $k = 1$ for illustration, if a main effect is the true response it is harder to eliminate a j-factor ($j > 1$) interaction than it would be to eliminate the main effect were the true response that j-factor interaction.

8. General response functions

A feature of interactions that greatly facilitates the computations is that $p(z|\delta)$, as z ranges over all 2^{m+1} values, assumes only 2 values (Theorem 1) with the result that any E (equation (3.2)) only contains 2 logarithmic terms. For a general response many more values will arise and the calculation of E's, and hence Δ's, is formidable. Some understanding of what happens with a general response can be obtained by calculating a generalized Kullback-Leibler number

$$\Delta(\delta : \delta_1, \delta_2) = \sum_z p(z|\delta) \cdot \log \frac{p(z|\delta_1)}{p(z|\delta_2)} \qquad (8.1)$$

$$= E(\delta : \delta_1) - E(\delta : \delta_2)$$

where δ_1 and δ_2 are still interactions but δ is general. This is the expected change in log-odds from one unit when comparing two interactions but when the true response is δ. Even in complete factorial experiments (section 1) it is usual to do the analysis in terms of the interactions even though the overall effect is not a pure

interaction. There any effect is a linear combination of interactions: the same property holds here. Hence a study of interactions with a general response is not inappropriate.

We now proceed to evaluate $E(\delta : \delta_1)$ where δ is arbitrary and δ_1 is an interaction. There are two ways to do this: either by an extension of the argument already used when δ is an interaction or by using the theory of vector spaces in which δ is expressed as a linear combination of interactions. The second method is simpler and has the added advantage that the structure of the situation is more clearly revealed. The reader not interested in mathematical details may proceed to the statement of Theorem 3 though equation (8.2) has to be understood in order to appreciate the role of the w's in the statement of that theorem.

To express any response function as a vector in 2^m-space write out the 2^m possible ξ's in any fixed order. The Table gives such a list for $m = 3$. The values $\eta(\xi)$ form a string of 2^m pluses and minuses and provide a complete description of δ: we shall denote this vector by δ. Such a representation of a δ is given in the fourth column of the Table. For any two different δ's consider the scalar product $\delta_1^T\delta_2$. The contribution to this will be $+1$ if δ_1 and δ_2 agree at a particular place and -1 if not. Let N denote the number of agreements between δ_1 and δ_2 so that 2^m-N is the distance between them. Then

$$\delta_1^T\delta_2 = N - (2^m-N) = 2(N-2^{m-1})$$

and we write

$$2^{-m}\delta_1^T\delta_2 = 2^{-(m-1)}(N-2^{m-1}) = w \quad . \qquad (8.2)$$

If $\delta_1 = \delta_2$ then $2^{-m}\delta^T\delta = 1$: if δ_1 and δ_2 are complementary $2^{-m}\delta^T\delta = -1$. If δ_1 and δ_2 are any two different and non-complementary interactions, $N = 2^{m-1}$ and $w = 0$. It follows that the 2^m odd interactions form an orthogonal basis for the vector representation of δ's. (The same holds for the even interactions.) We may therefore write $\delta = \sum a_i\delta_i$ where δ is arbitrary, $\{\delta_i\}$ is the set of odd interactions and $\{a_i\}$ is a set of numbers to be found. For interaction δ_s, $\delta^T\delta_s = \sum a_i\delta_i\delta_s = a_s 2^m$ and consequently from (8.2) $w_s = a_s$. We therefore have the result that any response can be written as a linear combination of odd interactions with weights w equal to $2^{-(m-1)}N_s - 1$ where N_s is the number of agreements between δ and the interaction δ_s: equation (8.2).

Consider the 2^m vectors of all odd interactions to form a square matrix. Then the number of minuses in every <u>row</u> of this matrix will be 2^{m-1} except for one row in which it will be 2^m. (The Table again illustrates for $m = 3$.) This exceptional row will be that corresponding to the ξ for which every ξ_i is -. If δ has also - in this row (if not, then use the representation in terms of even interactions) the total number of agreements between δ and all odd interactions will be $\sum N_s = 2^m + (2^m-1)2^{m-1}$ from which it follows that $\sum w_s = 1$.

Finally apply the transformation $\delta' = \frac{1}{2} + (p - \frac{1}{2})\delta$ to any δ which turns each $-$ into $1 - p = q$ and each $+$ into p. Then

$$\delta' = \frac{1}{2} + (p - \frac{1}{2})\delta = \frac{1}{2} \sum w_s + (p - \frac{1}{2}) \sum w_s \delta_s = \sum w_s \delta'_s \ ,$$

since $\sum w_s = 1$, and the representation is the same after transformation.

This last result enables $E(\delta : \delta_1)$ to be calculated rather easily. From (5.1)

$$p(z \mid \delta) = 2^{-m} \sum_{\zeta \in z(\delta)} p(z \mid \zeta) = 2^{-m} \sum_{\xi} p(x \mid \xi) p(y \mid \delta(\xi))$$

but from the result just established

$$p(y \mid \delta(\xi)) = \sum_s w_s \, p(y \mid \delta_s(\xi))$$

so that

$$\begin{aligned} p(z \mid \delta) &= 2^{-m} \sum_{\xi} p(x \mid \xi) \sum_s w_s \, p(y \mid \delta_s(\xi)) \\ &= 2^{-m} \sum_s w_s \sum_{\xi} p(x \mid \xi) p(y \mid \delta_s(\xi)) \\ &= \sum w_s \, p(z \mid \delta_s) \ . \end{aligned}$$

In words, the representation of δ in terms of interactions is also the representation of $p(z \mid \delta)$ in terms of $p(z \mid \delta_s)$. (Note that although $\sum w_s = 1$, the w's are not necessarily positive so that this is <u>not</u> a straightforward probability result.) It immediately follows from (8.1) that

$$E(\delta : \delta_1) = \sum_s w_s \, E(\delta_s : \delta_1)$$

and the $E(\delta_s : \delta_1)$-values are known from lemmas 2 and 3. Hence, again using $\sum w_s = 1$

$$E(\delta : \delta_1) = w_1 I(\alpha_{i+1}) + (1 - w_1)J(\alpha_{i+1}) - m \log 2$$

where i is the order of the interaction δ_1.

Theorem 3. For any response δ and for different and non-complementary interactions δ_s and δ_t of orders s and t respectively the generalized Kullback-Leibler number $\Delta(\delta : \delta_s, \delta_t)$ is

$$w_s I(\alpha_{s+1}) + (1 - w_s)J(\alpha_{s+1}) - w_t I(\alpha_{t+1}) - (1 - w_t)J(\alpha_{t+1}) \quad . \quad (8.3)$$

Here N_s is the number of agreements between δ and δ_s and $w_s = 2^{-(m-1)}(N_s - 2^{m-1})$.

The result follows directly from the evaluation of $E(\delta : \delta_s)$ and $E(\delta : \delta_t)$.

If $\delta = \delta_s$, $w_s = 1$ and $w_t = 0$ and we have Theorem 2, the special case of that Theorem being excluded.

Suppose $R(\delta_t)$ is not contained in $R(\delta)$, then in the representation of δ, δ_t will not appear and $w_t = 0$. If $R(\delta_t)$ is contained in $R(\delta)$ then by the relevancy principle we may confine our attention to the factors in $R(\delta)$. If these are k in number $w_t = 2^{-(k-1)}(N_t - 2^{k-1})$ where N_t refers only to agreements in $R(\delta)$.

To understand what is happening, suppose δ is any response with $R(\delta)$ containing $k(\leq m)$ factors. If two interactions not in $R(\delta)$ are compared (8.3) reduces to $J(\alpha_{s+1}) - J(\alpha_{t+1})$ which has the sign of $(s-t)$ so that a lower order interaction will appear less probable than a higher order one. If two interactions of the same order,

$s = t$, are compared, one within $R(\delta)$ and one not, (8.3) reduces to $w_s\{I(\alpha_{s+1}) - J(\alpha_{s+1})\}$ which is positive. Hence an interaction involving the k relevant factors will appear more probable than one of the same order involving some irrelevant factors. Of interactions within $R(\delta)$ of a given order that with the maximum weight will predominate. It is not possible to make general statements about interactions of different orders within $R(\delta)$ since the difference between $I(\alpha_{s+1})$ and $I(\alpha_{t+1})$ (and between the J's) depends on p whereas between w_s and w_t depends solely on δ and not p.

The Table gives an example with 3 factors. The first 3 columns are the possible ξ_i's making up the 8 possible ξ. The next column gives η for a δ in which the response is + if ξ_1 is + and, in addition, either ξ_2 or ξ_3, or both, are +: so $k = m = 3$. The remaining columns give η for the 8 odd interactions including the zero-factor interaction that always has a negative response. Under these last eight columns are listed the N's, the numbers of agreements between δ and the interaction of that column. In 3 cases $N < 2^{k-1} = 4$ so the number $2^{k-1} - N$ is listed below for the complementary even interaction. The resulting N's all exceed 4 and the w's all exceed 0. (This procedure of switching to complementary interactions is quite general.) The next row lists the w's and the final row provides the values of $E(\delta : \delta_s)$ for each interaction δ_s. The generalized Kullback-Leibler numbers are differences of E's and are not listed.

Consider the main effect of ξ_1 in comparison with that of ξ_2 (or ξ_3). The difference of the E's is -.59- (-.84) = +.25 and so ξ_1 will appear more probable than the other two, and will dominate all other interactions since it has the largest E. The next most important interaction is the 3-factor one which will slightly dominate the 3 2-factor ones, and all will dominate the ξ_2 and ξ_3 main effects.

Another interesting response (not tabulated) is that in which η is + only when all of $\xi_1, \xi_2, \ldots \xi_k$ are +. In medical language the disease is only present if k (true) symptoms are present. Such a δ differs from a zero-factor interaction in only one place and hence has $2^{k-1} + 1$ agreements with any other interaction (allowing for complements) giving $w_s = 2^{-(k-1)}$ for all interactions involving some or all of the $\xi_1, \xi_2, \ldots \xi_k$. It is not difficult to show that $w\, I(\alpha_{s+1}) + (1 - w)J(\alpha_{s+1})$ increases in s if $w < \frac{1}{2}$ so that (8.3) shows that the k-factor interaction of $\xi_1, \xi_2, \ldots \xi_k$ will dominate all other interactions in $R(\delta)$ and, by the general result, will dominate all k-factor interactions.

9. Conclusions

The main lesson to be learnt from this analysis is that there are situations in which the increasing complexity that inevitably arises from an increase in the number of factors gives rise to a need for more observations to unravel the complexity; and that this is in contrast to other familiar statistical situations in which the observational

explosion does not take place. The practical consequence of this is that statisticians should be wary of data analyses involving many factors because, if a model like that studied here is resonable, there just may not be enough observations to permit a satisfactory analysis, so that any data analysis must be a waste of time. Statisticians are often appealled to "to make sense out of this data": they should resist the temptation to do so without first checking that the extraction is possible. Rather their talents should be directed toward sensible designs and scientists encouraged to do planned experimentation rather than idle data collection.

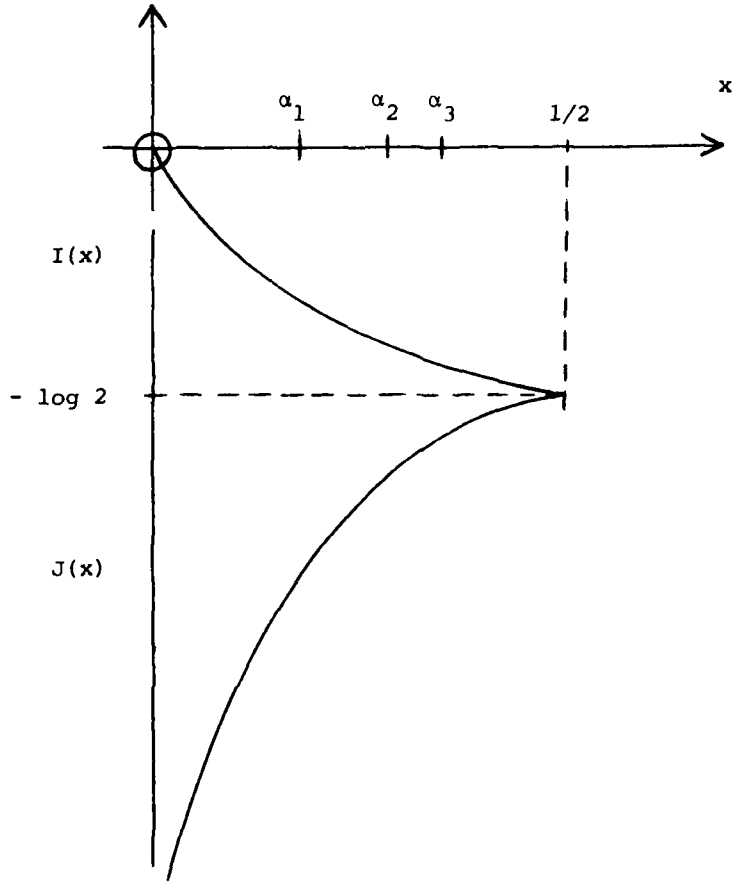
The analysis has also shown the usefulness of the Kullback-Leibler distance as a way of separating the possibilities and providing discrimination between them. A useful way of understanding what is happening with this model is to think of the "space" of possible δ's packed so tight that the "distances" between δ's become small. In contrast, the m-factor situation mentioned in the introduction has δ's much more loosely packed so that discrimination is easier.

Table for three factors ($m = 3$)

ξ			δ	1-factor			2-factor			3
				ξ_1	ξ_2	ξ_3	$\xi_1\xi_2$	$\xi_2\xi_3$	$\xi_3\xi_1$	$\xi_1\xi_2\xi_3$
-	-	-	-	-	-	-	-	-	-	-
-	-	+	-	-	-	+	-	+	+	+
-	+	-	-	-	+	-	+	+	-	+
+	-	-	-	+	-	-	+	-	+	+
+	+	-	+	+	+	-	-	+	+	-
+	-	+	+	+	-	+	+	+	-	-
-	+	+	-	-	+	+	+	-	+	-
+	+	+	+	+	+	+	-	-	-	+
			N:	7	5	5	3	5	3	3
							(5)		(5)	(5)
			w:	3/4	1/4	1/4	1/4	1/4	1/4	1/4
			E:	-.59	-.84		-.77			-.74

N is the number of agreements between δ and the interaction of that column: when this is less than 2^{m-1} (here 4), N for the complementary interaction is given in brackets. $w = 2^{-(m-1)}N - 1$ (equation (8.2)) with $N > 2^{m-1}$. $E = E(\delta, \delta_k)$ for interaction δ_k (equation (3.2)).

Figure



$I(x) = x \log x + (1-x)\log(1-x)$: $J(x) = \frac{1}{2} \log x(1-x)$:

$$\alpha_n = \frac{1}{2} - \frac{1}{2}(q-p)^n.$$

REFERENCES

LIN, SHU (1970). An introduction to error-correcting codes. Englewood Cliffs: Prentice-Hall Inc.

DVL/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2246	2. GOVT ACCESSION NO. AD-A103 877	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Relationship Between the Number of Factors and Size of an Experiment		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) D. V. Lindley		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE July 1981 13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

diagnosis; error-correcting codes; errors in variables; experimental design; interactions; Kullback-Leibler numbers; non-parametric methods

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

If the number of factors in an experiment is increased, does it necessarily follow that the size of the experiment must increase to achieve a satisfactory analysis? In some common situations the answer is No. The present paper discusses a model suggested by medical diagnostic problems in which the answer is Yes: indeed, the increase in size is exponentially fast. The conclusion is drawn that statisticians should be cautious before embarking on the study of data with large numbers of factors because the data may be inadequate for a sensible analysis. The basic, mathematical tool is the Kullback-Leibler number which

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (continued)

measures the discrimination between the possibilities. Calculation of these numbers uses interactions, forming a basis for all the effects that might occur.

END

DATE
FILMED
10-81

DTIC